SALIENT-MOTION-HEURISTIC SCHEME FOR FAST 3D OPTICAL FLOW ESTIMATION USING RGB-D DATA

Can Wang, Hong Liu

Engineering Lab on Intelligent Perception for Internet of Things(ELIP),
Shenzhen Graduate School, Peking University, China

ABSTRACT

Optical flow is widely used for describing motion cues in the scene, but limited by slow estimating speed and illumination sensitivity. To handle both problems, this paper focuses on improving speed and accuracy of optical flow using RGB-D data and enhancing its robustness on motion description via a fusing depth flow which is obtained only using depth data. First, salient motion regions (SMRs) are detected between depth frames which have good character on motion description for they all locate on moving objects. Then, depth flow is calculated to describe 3D motion for each SMR and directs fast orientation region growing on depth map. Thus larger motion regions are grown, and region-based optical flow estimation is conducted on grown regions. Estimation error is reduced and noise is inhibited due to depth constraints. Finally, a fusion scheme is adopted which combines depth flow and optical flow for better 3D motion description in the scene. Experiments on a RGB-D video data sets recorded in various complex scenes demonstrate the improved speed and robustness of the proposed method.

Index Terms— 3D optical flow, depth flow, RGB-D

1. INTRODUCTION

Optical flow is one of the traditional techniques in carrying out motion analysis. It measures the apparent velocity pattern of moving structures in an image sequence [1]. However, 2D optical flow is just projection of 3D motion of the world on 2D image plane and cannot reflect all motion cues. Computing 3D motion of a scene is a basic task in computer vision that has been approached in a wide variety of ways. Structure-from-motion [2][3] method is used to compute 3D scene structure and relative motion from a single monocular video sequence. However, without strong enough priori assumptions about the scene, general non-rigid motion cannot be estimated from a single camera. Another common approach to recover 3D motion is motion-stereo which uses multiple cameras and combine stereo and motion [4]. However, nearly all motion-stereo algorithms assume the scene is rigid and require fully calibrated cameras. Recently, with rapid development of range sensors many researches introduced depth data to motion analysis. M.B. Hotle et.al. use depth as well as intensity images captured by a SR4000 camera to extract 3D optical flow for motion recognition [5]. However, their 3D optical flow vectors is obtained directly via annotating each point on depth map with 2D optical flow vector, resulting in higher computational complexity than dense 2D optical flow. C. Wang et.al. use depth data captured by PrimeSense sensor for depth motion detection and multi-objects tracking [6]. However, only depth data cannot promise dense and subtle motion description compared with optical flow. B.B. Ni et.al. combine RGB and depth data captured by Kinect sensor to describe motion for activity recognition [7]. But depth information is not fully exploited and it only assists depth-level division of STIP descriptors.

Directly extracting dense 3D optical flow using state-of-the-art optical flow methods [8][9] are far more computation consuming and cannot fully utilize depth cues. Indeed, the basic function of 3D optical flow is to describe motion, so it is reasonable to first adopt a fast way to localize salient motion regions in the scene. Motivated by this, a salient-motion-heuristic scheme for fast 3D optical flow estimation is proposed. First, regions with salient depth motion are extracted only using depth data. Then ability of depth on motion description is further exploited, and a kind of region-based velocity named depth flow is calculated. Directed by depth flow, a fast orientation region growth is conducted for larger motion regions and based on which dense optical flow is estimated finally. Our contribution lies in three aspects: One is the improved speed for pre-detected salient motion localizes and lessens flow estimation regions. The second one is the inhibition of noise commonly caused by illumination change yet to which depth cue is robust. The third is the fusion scheme of optical and depth flow makes both cues complementary for more robust motion description. Contrast experiments show the proposed method not only reduces computational cost but also reduces noise interference and multi-tracking experiments under various complex situations verify robustness of the fusion scheme.

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JC201005280682A, CXC201104210010A).

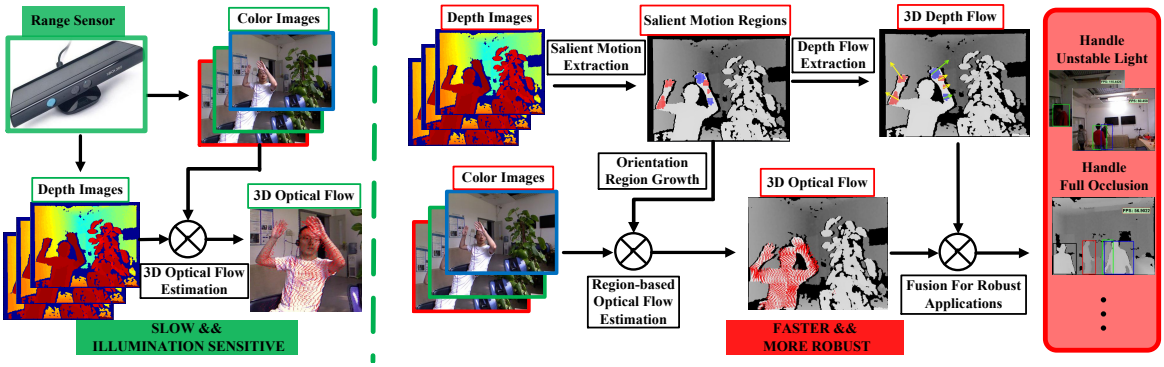


Fig. 1: A brief illustration of previous 3D optical flow estimation methods (left part) and the proposed method (right part).

2. SALIENT MOTION HEURISTIC 3D OPTICAL FLOW ESTIMATION

2.1. Salient Motion Detection

Different from intensity images, points in depth images in essence represent 3D positions in real world, thus depth images sequence essentially represents the variation of these positions. As shown in Fig.2, points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ both change their 3D positions during time $t-1$ to $t+1$, so their depth values also change. In our previous work [6][10], the concept 'positive motion' is defined to indicate a specific kind of significant depth change in consecutive depth images. B.B. Ni et.al. also present the similar concept backward motion and forward motion to describe depth motion [7] .

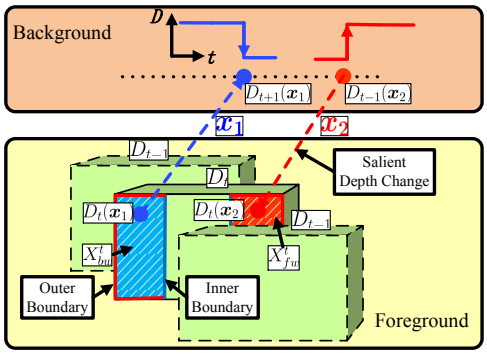


Fig. 2: Illustration of salient depth change, salient motion region X^t_{fw} and X^t_{bw}, and their inner and outer boundaries. Depth frames in $t-1$ and $t+1$ are combined together to extract motion cues in current frame t.

Under the assumption of smooth motion, the significant depth change of a point $\boldsymbol{x}_i$ normally indicates that position of $\boldsymbol{x}_i$ changes from one object to another and there exists a considerable depth difference between them. It comes to a natural idea that we may utilize points which exhibit salient depth change between consecrative frames D_{t-1} and D_{t+1} to describe the motion in the current frame D_t. Based on this intuition, positive motion point (PMP) is proposed. A great nature of PMP is that ideally they all locate on moving objects, thus brings bonus to describe and analyze motion cues. This specific salient motion point is defined as:

$$\begin{aligned} X^t_{pm} &= X^t_{fw} \bigcup X^t_{bw} \quad \text{where} \\ X^t_{fw} &= \{\boldsymbol{x} | D_t(\boldsymbol{x}) - D_{t-1}(\boldsymbol{x}) > \tau_{pm}, D_{t-1}(\boldsymbol{x}) > \tau_{ur}\} \\ X^t_{bw} &= \{\boldsymbol{x} | D_t(\boldsymbol{x}) - D_{t+1}(\boldsymbol{x}) > \tau_{pm}, D_{t+1}(\boldsymbol{x}) > \tau_{ur}\} \end{aligned} \tag{1}$$

At a given coordinate $\boldsymbol{x}$, $D_t(\boldsymbol{x})$ is pixel value of depth map D at time t. τ_{pm} is a threshold indicating whether there is a salient depth change in $\boldsymbol{x}_i$. τ_{pm} is set to define the salient level of depth change. τ_{ur} is set to remove salient depth change of unstable regions [6], which is normally caused by hardware drawbacks of the range sensor [7][11] ,such as smooth surface and transparent objects.

2.2. Orientation Region Growing and 3D Depth Flow

Suppose there are K connected regions in points set X^t_{pm} on depth map. For each region $X^{(k)}_{pm}$, its boundary point set can be divided into inner boundary points set and outer boundary points set via gradient analysis on depth map, termed as $X^{(k)}_{ib}$ and $X^{(k)}_{ob}$ respectively. A brief illustration is given in Fig.2. Normally, the outer boundary is the edge of a salient motion region $X^{(k)}_{pm}$ in current frame, and the inner boundary is the edge of the same region in the consecutive frame $\hat{D}_t$, where $\hat{D}_t$ is defined as:

$$\hat{D}_t = \begin{cases} D_{t-1} & \text{for } X^{(k)}_{pm} \in X^t_{fw} \\ D_{t+1} & \text{for } X^{(k)}_{pm} \in X^t_{bw} \end{cases} \tag{2}$$

Given inner and outer boundaries, 2D velocity of region $X^{(k)}_{pm}$ can be calculated by:

$$\boldsymbol{v}^{(k)} = \delta \cdot \left(\overline{X^{(k)}_{ib}} - \overline{X^{(k)}_{ob}} \right) \tag{3}$$

where $\overline{X}$ indicates center of point set X and δ is defined as:

$$\delta = \begin{cases} 1 & \text{for } X^{(k)}_{pm} \in X^t_{fw} \\ -1 & \text{for } X^{(k)}_{pm} \in X^t_{bw} \end{cases} \tag{4}$$

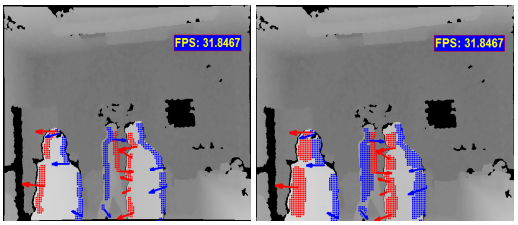


Fig. 3: Illustration of salient motion regions before and after ORG with α=0.5 and their corresponding 3D depth flows.

Then 2D velocity $\boldsymbol{v}^{(k)}$ of kth salient motion region can direct us to perform orientation region growth (ORG) to get larger motion region on moving object's surface. In order to boost speed of region growth, besides orientation guidance, state matrix $\hat{\boldsymbol{S}}$ is defined to avoid repeated growth on the same location for two or more regions. $\hat{\boldsymbol{S}}(\boldsymbol{x})$ is set to 1 if $\boldsymbol{x}$ is grown. ORG is conducted on depth map, formulated as:

$$\begin{aligned} X_{rg}^{(k)} &= \bigcup_{\boldsymbol{x}_j \in X_{ib}^{(k)}} \bigcup_{t_0 \in [0,\alpha]} \left(\boldsymbol{x}_j + t_0(-\delta \cdot \boldsymbol{v}^{(k)}) \right) \\ s.t. \quad & \forall \boldsymbol{x} \in X_{rg}^{(k)}, D(\boldsymbol{x}) \in [d_l, d_h]; \quad \hat{\boldsymbol{S}}(\boldsymbol{x}) = 0 \end{aligned} \tag{5}$$

where α is set to control growing scale. Here, $[d_l, d_h]$ is the depth interval constraints for growing. ORG actually can be summarized as a process of region growth on depth map, starting from inner boundary of a salient motion region, along with orientation (when δ equals -1) or inverse orientation (when δ equals 1) of the region's 2d velocity (given in Eq.(3)).

Given ORG region $X_{rg}^{(k)}$ of kth salient motion region, its velocity along depth axis can be obtained by calculating average depth change between two consecrative frames:

$$v_d^{(k)} = \delta \cdot \left(\overline{D_t\big(X_{pm}^{(k)}\big)} - \overline{\hat{D}_t\big(X_{pm}^{(k)}\big)} \right) \tag{6}$$

Where $\overline{D(X)}$ is used to indicate mean depth value of a point set. If $X_{pm}^{(k)}$ belongs to X_{fw}^t, $\hat{D}_t$ corresponds to D_{t-1} and δ equals 1, and if $X_{pm}^{(k)}$ belongs to X_{bw}^t, $\hat{D}_t$ corresponds to D_{t+1}, since the time sequence inverse, δ should equal to -1, this makes the velocity always describe motion from current frame to next frame. Combined 2d velocity $\boldsymbol{v}^{(k)}$ with depth velocity $v_d^{(k)}$, 3D depth flow of the kth salient motion region is proposed to describe its 3D motion, formulated as:

$$\boldsymbol{f}_d^{(k)} = \big(T(\boldsymbol{v}^{(k)}),\ T(v_d^{(k)})\big) \tag{7}$$

Where T indicates heterogeneous data conversion because $\boldsymbol{v}^{(k)}$ and $v_d^{(k)}$ have different length unit (image coordinates and range axis). To simplify symbols, we use T for all heterogeneous data conversion in this article.

2.3. Fusion Scheme of 3D Optical Flow and Depth Flow

After ORG, salient motion regions are fused to larger ones by connectivity on depth map. Suppose a fused salient motion region is termed as X_{sm}, we proposed a region-based optical flow method. First, a parallel implementation of LucasKanade (LK) method [12] is utilized to estimate 2D dense optical flow between intensity image patch $I(X_{sm})$ and $\hat{I}(\hat{X}_{sm})$ in consecutive frame $\hat{I}$, where region $\hat{X}_{sm}$ is the largest points set which satisfies:

$$\begin{aligned} &\text{for} \quad \forall \boldsymbol{x}_i \in \hat{X}_{sm}, \quad \exists \hat{\boldsymbol{x}}_i \in X_{sm} \\ &\sqrt{T(\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|)^2 + T(|\hat{D}(\hat{\boldsymbol{x}}_i) - D(\boldsymbol{x}_i)|)^2} \le \tau_v \end{aligned} \tag{8}$$

τ_v is upper bound of 3D optical flow length. This 3D Euclidean distance constraint reduces background noise and improve accuracy of optical flow in salient motion region. Finally, 2D optical flow is estimated on all salient motion regions. Suppose 2D optical flow in $\boldsymbol{x}_i$ is $\big(v_x(\boldsymbol{x}_i), v_y(\boldsymbol{x}_i)\big)$, its depth component $v_z(\boldsymbol{x}_i)$ is calculated by:

$$v_z(\boldsymbol{x}_i) = \delta \cdot \big(\hat{D}_t\big(\boldsymbol{x}_i + \delta \cdot \big(v_x(\boldsymbol{x}_i), v_y(\boldsymbol{x}_i)\big)\big) - D_t(\boldsymbol{x}_i)\big) \tag{9}$$

The final salient-motion-based 3D optical flow in any point $\boldsymbol{x}_i$ is given by:

$$\boldsymbol{f}(\boldsymbol{x}_i) = \big(T\big(v_x(\boldsymbol{x}_i)\big),\ T\big(v_y(\boldsymbol{x}_i)\big),\ T\big(v_z(\boldsymbol{x}_i)\big)\big) \tag{10}$$

Moreover, a fusion scheme is adopted to fuse optical flow $\boldsymbol{F}$ and depth flow $\boldsymbol{F}_d$ in tracking applications, which makes two cues complementary for each other when one is unreliable. Thus, motion cues can be more accurately described. For jth target during tracking, its current motion velocity $\boldsymbol{v}^j$ is given by:

$$\begin{aligned} r = \frac{1}{\sigma(\frac{\boldsymbol{f}^j}{\|\boldsymbol{f}^j\|})} \quad r_d = \frac{1}{\sigma(\frac{\boldsymbol{f}_d^j}{\|\boldsymbol{f}_d^h\|})} \\ \boldsymbol{v}^j = \frac{r}{r + r_d} \overline{\boldsymbol{f}^j} + \frac{r_d}{r + r_d} \overline{\boldsymbol{f}_d^j} \end{aligned} \tag{11}$$

Where $\boldsymbol{f}^j$ and $\boldsymbol{f}_d^j$ are optical flow and depth flow on jth target area. r and r_d indicate reliability of optical flow and depth flow respectively. δ is variance and $\overline{\boldsymbol{f}_i}$ indicates mean of flow vector set $\boldsymbol{f}_i$. As the tracking term is not the key point of the article, it is directly verified in experiments section.

3. EXPERIMENTS AND DISCUSSIONS

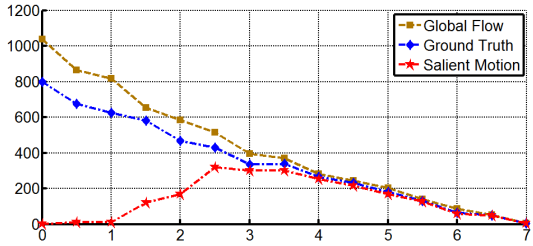
To demonstrate effectiveness of the proposed method, three groups of experiments are conducted on a RGB-D video data set recorded via Kinect. All experiments were conducted on a Pentimum $i5-2410M$ 2.3 GHZ PC with 2.0 Gb RAM.

The first group of experiments verify the speed improvement of 3D optical flow estimation based on salient motion, compared with 3D version of state-of-the-art dense optical flow methods. The experiments is tested under various video resolutions and processing frame rates (FPS) are given in Table 1. SM-LK, SM-HS, SM-FFC are our salient-motion-based optical flow methods corresponding to LK, HS and FC-C (fast cross correlation) methods implemented in [12][13] respectively.

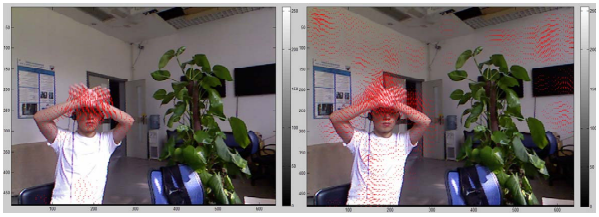
Table 1: Optical flow extraction speed (in FPS) comparison.

Resolutions	80*60	160*120	320*240	640*480
LK [[12]	76.9	45	10.1	2.3
SM-LK	**113.2**	**86.5**	**35.2**	**10.5**
HS [[12]	4.1	1.3	0.2	0.05
SM-HS	**30.6**	**17.8**	**8.3**	**3.2**
FCC [[13]	12.3	7.2	2.1	0.2
SM-FCC	**63.8**	**21.7**	**8.3**	**1.5**
CC [[12]	0.83	0.19	0.05	0.01

The second group of experiments verify accuracy of optical flow via the proposed method. The statistics of optical flow vectors of the proposed method, ground truth and LK flow are compared which demonstrates the proposed method is more accurate in extracting salient motion cues and is more robust to background noise. As shown in Fig.4 (a), average number distribution of 3D flow vectors along different vector length per frame is given. The number of globally estimated flow is higher than ground truth because of false extraction in noise regions due to unstable imaging or tiny motion. On the contrary, the proposed method can extract most flow vectors with longer length and remove tiny motion which is often caused by noise in real cases. As shown in Fig.4 (b), lots of noisy flow vectors on background are removed in the left compared with the right one.



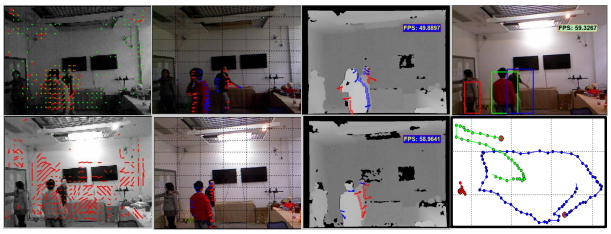
(a) Number distribution of 3D flow vectors along different vector length per frame. In top-right legend, ground truth is extracted via LK on manually-labeled salient motion region. Global flow is obtained via LK. Salient motion is via the proposed method.



(b) An example of flow distribution in a clutter scene with a man waving hands. The right is result via the proposed method. The left is via LK. It is easy to see the proposed method extract salient motion on moving body parts but inhibit noise on background caused by unstable lighting.

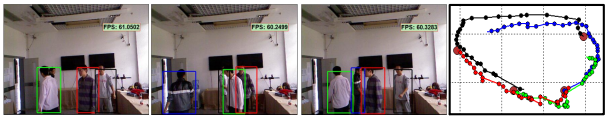
Fig. 4: Optical flow extraction accuracy comparison.

The third group of experiments verify the robustness of the fusion scheme which combines 3D depth flow and 3D optical flow in describing motion. Several tracking experiments are conducted in various complex scenes.



(a) The first column shows classic optical flow when illumination suddenly change. The second column and third column is optical flow and depth flow extracted via the proposed method under same illumination conditions. The forth column is targets' labels and trajectories.

(b) Multi-tracking under similar foreground and background, and bad lighting conditions.



(c) Multi-tracking under severe occlusions.

Fig. 5: Robustness verification of fusion scheme combining optical flow and depth flow for motion description.

4. RELATION TO PRIOR WORK AND CONCLUSIONS

This paper focuses on improving speed and accuracy of optical flow and enhancing its robustness on motion description. M.B. Holte et.al. directly transform dense 2D optical flow implemented via hierarchical LK to 3D version flow, simply using depth value of each point on depth map [5] which cannot bring speed improvement. Thus one of our motivation is to estimate dense optical flow based on pre-detected salient motion cues, for the basic function of optical flow is representing salient motion in the scene. For salient motion detection only using depth data which is robust to illumination change, so our method works fine with bad illumination. As other works which also use depth-color fusion scheme for various applications [7][10][14][15] , but all of them emphasis on motion description using RGB data and only utilize depth to facilitate the former one. To the contrast, this work deeply exploit the ability of depth on motion description and introduced depth flow which can be fast extracted only via depth. Then, our fusion scheme combines sparse depth flow and dense optical flow together for more robust motion representation.

In conclusion, contribution of this work lies in speed improvement, accuracy of 3D optical flow estimation and robustness motion description via the complementary of depth can optical flow. Contrast experiments show the proposed method not only reduces computational cost but also reduces noise interference and multi-tracking experiments under various complex situations verify robustness of the fusion scheme. In future work, high level motion description based on 3D flow will be further researched.

5. REFERENCES

[1] A. Becciu , H. van Assen , L. Florack , S. Kozerke , V. Roode and B. ter Haar Romeny "A multi-scale feature based optic flow method for 3D cardiac motion estimation", Scale Space Variant Methods Computer Vision, 2009, pp.588-599. 1

[2] J.P. Costeira and T. Kanade, "A multibody factorization method for independently moving Objects", International Journal of Computer Vision, IJCV 1998, pp.159-179. 1

[3] S. Avidan and A. Shashua., "Non-rigid parallax for 3D linear motion". IEEE Conference on Computer Vision and Pattern Recognition, CVPR 1998, pp.62-66. 1

[4] G.S. Young and R. Chellappa., "3-D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness". IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI 1999, pp.735-759. 1

[5] M.B. Holte, T.B. Moeslund and P. Fihl."View-invariant gesture recognition using 3D optical flow and harmonic motion context". Computer Vision Image Understanding, CVIU 2010, pp.1353-1361. 1, 4

[6] C. Wang and H. liu, "A Reliable Moving Human Detection and Tracking Method based on Accurate Objects Segmentation using Range Sensor",IEEE International Conference on Multi-sensor Fusion and Information Integration, MFI 2012, Hamburg, German, 13-15 September, pp.330-335. 1, 2

[7] Bingbing Ni, Gang Wang and Pierre Moulin, "RGBD-HuDaAct: A Color-Depth Video Database For Human Daily Activity Recognition", IEEE workshop on Consumer Depth Cameras for Computer Vision, in conjunction with ICCV 2011. 6-13 November pp.1147-1153. 1, 2, 4

[8] B. Horn and B. Schunck , "Determining Optical Flow", Artificial Intelligence ,1981, vol. 17, pp.185-203.

[9] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision". Proceedings of Imaging Understanding Workshop, 1981, pp.121-130. 1

[10] C. Wang and H. liu, "Robust Visual Tracking based on Adaptive Depth-Color-Cue Integration using Range Sensor",IEEE International Conference on Multi-sensor Fusion and Information Integration, MFI 2012, Hamburg, German, 13-15 September, pp.336-343. 1 2, 4

[11] Microsoft Corp. Redmond WA. "Kinect for Xbox 360". 2

[12] Piotr Dollar, Vincent Rabaud, Garrison Cottrell and Serge Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features", IEEE International Conference on Computer Vision VS-PETS 2005, Beijing, China, 15-16 October, pp.65-72. 3, 4

[13] Schindler, K.and van Gool, L. , "Action snippets: How many frames does human action recognition require?," IEEE Computer Vision and Pattern Recognition, CVPR 2008 , 23-28 June 2008, pp.1-8. 3, 4

[14] Spinello, Luciano and Arras, Kai O. , "People detection in RGB-D data", 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, 25-30 Septempber, pp.3838-3843. 4

[15] Bonnin, A., Borras, R.and Vitria, J. , "A cluster-based strategy for active learning of RGB-D object detectors", 2011 IEEE International Conference on Computer Vision Workshops, ICCVW 2011, 6-13 November, pp.1215-1220. 4